

Analysis of incomplete data using inverse probability weighting and doubly robust estimators

Stijn Vansteelandt¹, James Carpenter² and Michael G. Kenward²

¹ Ghent University, Ghent, Belgium

² London School of Hygiene and Tropical Medicine, London, U.K.

This article reviews inverse probability weighting methods and doubly robust estimation methods for the analysis of incomplete data sets. We first consider methods for estimating a population mean when the outcome is missing at random, in the sense that measured covariates can explain whether or not the outcome is observed. We then sketch the rationale of these methods and elaborate on their usefulness in the presence of influential inverse weights. We finally outline how to apply these methods in a variety of settings, such as for fitting regression models with incomplete outcomes or covariates, emphasizing the use of standard software programs.

Key words: doubly robust estimation; extrapolation; extreme weights; Horvitz-Thompson estimator; inverse probability weighting; missing data, multiple imputation.

1. Introduction

Missing data are often encountered in social science studies. They raise concern over standard analyses which are restricted to subjects with complete data, as these subjects may form an unrepresentative subgroup from whom biased conclusions may be obtained. The idea that this bias can be corrected by weighting each of these subjects'

observations by the inverse of the probability of observing complete data, has been around at least since Horvitz and Thompson formally introduced it in 1952 (Horvitz & Thompson, 1952). Nevertheless, for many years, the method of inverse probability weighting (IPW) gained little acceptance in the missing data literature because of its imprecision relative to more popular missing data methods, such as multiple imputation (Rubin, 1987). This has changed drastically over the past decade, since the seminal work of Robins, Rotnitzky and Zhao (1994), who demonstrated how the precision of IPW estimators could be greatly improved in a general regression context to the point where they become competitive with imputation estimators (Carpenter, Kenward & Vansteelandt, 2006). More recent work by Scharfstein, Robins and Rotnitzky (1999) and Robins and Rotnitzky (2001) has also contributed to this. These authors demonstrated that some IPW estimators possess a property of double robustness. Estimators that share this property are unbiased in large samples when either an imputation model or a model for the probability of complete data is correctly specified by the user, but not necessarily both. These estimators therefore enjoy greater robustness against model misspecification than both imputation and IPW estimators. Despite these advances, the practical usefulness of (doubly robust) IPW methods continues to be a matter of debate (see e.g. Kang & Schafer, 2008, and the subsequent discussions), partly because the literature on this topic is not easily accessible, and mostly because of concerns about the instability of IPW estimators in the presence of influential weights.

The goal of this article is to give an accessible introduction to inverse probability weighting methods and doubly robust estimation methods for the analysis of incomplete data. The key concepts are outlined in the second section where the focus is on estimating

a population mean. We contribute to the ongoing debate by elaborating on the usefulness of IPW and doubly robust estimators in settings where they tend to give results most distinct from imputation estimators (namely in the presence of influential inverse weights). We do this in the third section, both using large sample arguments and via simulation studies. The generality and flexibility of the inverse weighting idea is demonstrated in the fourth section in the context of fitting regression models with incomplete outcomes or covariates. This section may be skipped by the less technically minded reader upon first reading. Our emphasis throughout is on the intuition behind the methods and on the use of standard software. We end with a discussion of the relative advantages of the different estimation strategies.

2. Estimating a population mean from incomplete outcome data

2.1. Inverse probability weighting

Suppose that we have a study in which it is intended to collect outcome measurements Y_1, \dots, Y_n on a random sample of n subjects, but that these outcomes are missing for some subjects. Specifically, we let $R_i = 1$ denote that Y_i is observed and $R_i = 0$ denote that Y_i is missing. When missingness (i.e., whether or not the outcome is observed) is associated with prognostic factors X_i of the outcome, then the subjects with complete data form a selective subgroup and thus the sample average of their outcomes may systematically over/underestimate the population mean. This selection bias can be corrected when all prognostic factors X_i for missingness have been measured, in which case the data (Y_i, X_i) follow a so-called missing at random mechanism. This correction can be done by weighting each responder's data by the reciprocal of the probability π_i of

that subject having observed outcome data Y_i on the basis of his/her background characteristics X_i . In particular, having estimated the probabilities π_i – for instance by fitting a logistic regression model for the probability of observed data (i.e., $R_i = 1$), given the background characteristics X_i – we calculate

$$\hat{\mu}_{IPW} = \frac{\sum_{i=1}^n R_i Y_i / \pi_i}{\sum_{i=1}^n R_i / \pi_i}, \quad (1)$$

where $R_i Y_i$ is ‘ $R_i \times Y_i$ ’ and so is defined as zero whenever R_i is zero. The estimator $\hat{\mu}_{IPW}$ is called an inverse probability weighted (IPW) estimator. It is a variant of the Horvitz-Thompson estimator, which was introduced in the context of survey sampling in finite populations (Horvitz & Thompson, 1952). In practice, it is easily obtained with standard software via the following two-stage procedure:

1. Fit a logistic regression model for the probability of observing the outcome measurements (i.e., $R_i = 1$) as a function of prognostic factors X_i . Below, we refer to this model as the response model. Let π_i denote the fitted value for subject i .
2. Fit a linear model to the observed outcome measurements (with no predictors), using weighted least squares regression with weights $1/\pi_i$. The only estimated coefficient is the intercept, i.e. estimated mean response, which we denote by $\hat{\mu}_{IPW}$.

The estimator $\hat{\mu}_{IPW}$ is unbiased for the mean outcome in the study population, provided that the sample size is ‘sufficiently’ large. Intuitively, this is because the impact of inverse probability weighting is to ‘reconstruct’ a random sample from the intended

study population, by giving more weight to subjects when they are less likely (on the basis of their background characteristics X_i) to be observed. Specifically, subjects with a 50%, 25%, ... chance of observed outcome data are only half, a quarter, ... as frequently seen in the observed sample as in the study population. Thus the observed data for such subjects are weighted 2 ($=1/0.5$) times, 4 ($=1/0.25$) times, ... in the analysis to account both for themselves and for missing subjects with the same background characteristics in the population. The following somewhat extreme example, adapted from Carpenter, Kenward and Vansteelandt (2006), illustrates this. Suppose that the data for 9 subjects are as given in Table 1. Then the true outcome mean is 2. Suppose now that outcomes are missing for subjects with $R_i = 0$. Then the outcome mean for responders (i.e., those with $R_i = 1$) is $13/6$, biased. To correct this bias, the IPW estimator $\hat{\mu}_{IPW}$ requires first estimating the probability of observed outcome data for each subject on the basis of the measured covariate X , which explains the missingness. This probability is $1/3$ for subjects in group $X_i = A$ because only 1 in 3 subjects have recorded outcome data in that group, and likewise 1 and $2/3$ for subjects with $X_i = B$ or C , respectively. The single outcome that was recorded for subjects with $X_i = A$ is thus counted three times in the IPW estimator (see Equation 1): once to account for the subject with observed outcome in that group and twice more for the 2 subjects with missing data in that group. Note that the impact of this is to reconstruct the original measurements. Likewise, the observed outcome measurements for subjects with $X_i = B$ or C are each counted 1 and 1.5 times, respectively. More generally, the principle behind IPW estimators is illustrated in Figure 1 where the top left panel shows a simulated complete dataset. Now suppose some of the outcomes were made missing with probability $1-\pi_i$, depending on X_i . These outcomes are

shown by question marks. The top right panel shows all observed measurements weighted by $1/\pi_i$, where the circle's area is proportional to the weight. The IPW estimator is the weighted average of these measurements. It thus gives more weight to the observed outcomes to the left of the panel to account for the relatively greater number of missing outcomes in that region.

A more subtle intuition for the IPW estimator comes from noticing that inverse probability weighting removes the association between missingness and prognostic factors X_i , and thus makes the missingness non-selective. Indeed, the dependence of missingness on background characteristics is entirely explained by the missingness probabilities π_i , and thus completely disappears after inverse probability weighting (since R_i/π_i is on average 1 at all levels of the background characteristics X_i (see Table 1) and thus missingness R_i is not associated with X_i after inverse probability weighting). Finally, from a more theoretical perspective, the large-sample unbiasedness of the IPW estimator follows from (i) missingness having no residual association with the outcome after adjusting for X_i and (ii) R_i/π_i being 1 on average (e.g. Table 1). As the sample size increases, this is enough to ensure the IPW estimator $\hat{\mu}_{IPW}$ and the intended sample average $\frac{1}{n} \sum_{i=1}^n Y_i$ are the same in expectation.

Despite their theoretical validity and computational simplicity, a drawback of IPW estimators is that they can behave very badly in examples where a few individuals receive a very large weights (Robins, Rotnitzky & Zhao, 1995; Robins & Wang, 2000; Kang & Schafer, 2008; Robins et al., 2008). This is likely to happen when measured background characteristics are strongly predictive of missingness in the outcome. In view of this, we consider a number of alternative estimators.

Figure 1 about here.

2.2. Imputation

Imputation estimators (Kenward & Carpenter, 2007, or in more detail, Rubin, 1987) tend to be less variable than IPW estimators in the presence of extreme weights (Robins & Wang, 2000). In regression mean imputation, for instance, a model for the expected outcome given the background characteristics X_i is fitted to the responders, and then the missing values are imputed with fitted values $m(X_i)$ from this model. Next, the sample average from the imputed data set is calculated:

$$\hat{\mu}_{IMP} = \frac{1}{n} \sum_{i=1}^n R_i Y_i + (1 - R_i) m(X_i). \quad (2)$$


This estimation principle is visualized in Figure 1 (bottom left panel) where the missing outcomes are replaced with their expectations $m(X_i)$ on the basis of the measured background characteristics X_i . More generally, one may consider multiple imputation (Rubin, 1987; Kenward & Carpenter, 2007). This would lead to an equivalent estimator of the population mean as in Equation 2 in the case of an infinite number of imputations. Multiple imputation will therefore not be explicitly considered in this article, although the conclusions drawn for regression mean imputation will extend to it.

In view of later results, it is useful to note that when $m(X_i)$ is obtained as the fitted value from a generalized linear model fit, Equation 2 is identical to the sample average of the fitted values in all subjects, i.e.

$$\hat{\mu}_{IMP} = \frac{1}{n} \sum_{i=1}^n m(X_i). \quad (3)$$

This is so because the outcome and fitted values from a generalized linear model analysis have the same sample average.

Throughout, we will refer to a model for the expected outcome in the responders,

given the missingness predictors (e.g., X_i), as an .
 I M P U T A T I O N M O D E L

This is the model leading to the mean imputations $m(X_i)$. When this model is correctly specified, the regression mean imputation estimator $\hat{\mu}_{IMP}$ gives an unbiased estimate of the mean outcome in the study population. This is because when missingness is completely explained by X_i , the predicted value $m(X_i)$ – although obtained by fitting a regression model to the responders’ data – is also the correct expected outcome in nonresponders with background characteristics X_i .

2.3. Model misspecification and extrapolation

The reliance on correct specification of the imputation model is a salient feature of imputation estimators, in the same way that reliance on the correct specification of the response probability π_i is characteristic of IPW estimators. Ideally, the choice between these estimators should thus, at least partly, be based on which of these two models is more likely to be correctly specified: the imputation model or the response model. Throughout, we will characterize both models as ‘working models’ because they are not guaranteed to be correct. While standard goodness-of-fit techniques may be adopted for assessing the adequacy of these working models, such techniques may have limited success for detecting misspecification of the imputation model. This is because this model should appropriately predict the outcome in nonresponders and thus fit well in the region where the latter’s X -measurements are situated. When the covariate distribution

shows large separation between responders and nonresponders, conventional model checking (of the imputation model) cannot detect misspecification in that region. The imputation model, which is fitted to the responders, is then not reliable for imputing the outcome in nonresponders, as it may involve serious extrapolations beyond the range of the observed data (Tan, 2008). This is illustrated in Figure 1 (bottom left panel), which shows that the nonlinear trend of the outcome in X is difficult to detect on the basis of subjects with observed data alone. More generally, the imputation estimator under a linear prediction model may be heavily influenced by extreme extrapolations. This is not revealed by the output of multiple imputation software and is thus entirely implicit. These same limitations hold for maximum likelihood and mean score methods, which estimate parameters in the presence of missing data by averaging over the conditional (imputation) distribution of the missing data.

IPW estimators do not suffer directly from this extrapolation problem because they merely rely on a model for the probability of observing the outcome data, given background characteristics X , and this model is estimated from all the units in the data set. Instead, these estimators must deal with extreme weights which are likely to arise when responders and nonresponders are dissimilar in terms of background characteristics (X). This is reflected in large standard errors for the IPW estimators, arising because a few subjects are very influential in the analysis so minor changes in their data (hence weights) may non-trivially affect the results. This instability of IPW estimators at least partly explains the reason why these estimators, which are easy to compute, are not routinely used in practice. Note however that their instability is a proper reflection of the separation of the background characteristics between responders and nonresponders (Tan,

2008), and thus of the lack of information on the population mean in the data from the nonresponders. Thus it should not necessarily be regarded as a disadvantage of the approach, but rather as a consequence of its transparency in such settings. In particular, when there is complete separation in the distribution of background characteristics between responders and nonresponders, then IPW estimators will break down – i.e., nonresponders will have response probabilities π_i equaling zero (and thus $R_i/\pi_i = 0/0$ is not defined for these subjects). As such, they make clear that there is no basis for inferring the expected outcome in nonresponders. In contrast, maximum likelihood and (multiple) imputation estimators with comparatively small standard errors, but possibly large bias, will still be produced because they are based on this implicit extrapolation (see Murray & Findlay (1988) for an example).

2.4. Doubly robust estimators

Imputation and maximum likelihood estimators can be substantially more precise than IPW estimators when their respective working models are correctly specified, as they make more efficient use of all subjects' data. However, as explained in the previous section, assessing the working imputation model is more subtle and can be impossible in some situations. This raises the question as to which estimator to use in a given setting. Doubly-robust estimators (Robins & Rotnitzky, 2001; Davidian, Tsiatis & Leon, 2005; Carpenter, Kenward & Vansteelandt, 2006) overcome the need for choosing between the two alternative working models by maintaining validity when either one, but not necessarily both of them is correctly specified. One such estimator is obtained by replacing the expected value $m(X_i)$ in (3) by the fitted value $m^*(X_i)$ of a generalized linear

model analysis of outcome Y_i on background characteristics X_i , fitted to the responders using the weights $1/\pi_i$ (Robins et al., 2008). In particular, they can be obtained using standard software via the following three-stage approach:

1. Fit a logistic regression model for the probability of observing Y_i (i.e., $R_i = 1$) as a function of prognostic factors X_i , and let π_i denote the fitted value for subject i .
2. Fit a generalized linear model for the outcome of responders in function of prognostic factors X_i using weights $1/\pi_i$ and let $m^*(X_i)$ denote the fitted value for subject i .
3. Take the sample average of the fitted values $m^*(X_i)$ of both responders and nonresponders as an estimate of the outcome mean.

This estimation principle is visualized in Figure 1 (bottom right panel) where the missing outcomes are predicted on the basis of the measured background characteristics X_i , but where outcomes receive larger weight in calculating the mean response when they are more likely to be missing. By thus focusing more on regions of the covariate space where the nonresponders are located, these predictions may succeed better at predicting the outcome in the study population.

The doubly robust nature of the estimator (i.e., Equation 3 with $m(X_i)$ replaced by $m^*(X_i)$) can be understood from the following argument, which may be skipped by the less technically minded reader. First, when the imputation model is correctly specified, then misspecification of the weights $1/\pi_i$ does not hamper the validity of the estimators $m^*(X_i)$, and hence of the doubly robust estimator. This is because the weights $1/\pi_i$ only depend on X_i and the model residuals have mean zero for each X_i . Second, when the response model is correctly specified, then misspecification of the imputation model does

not affect the validity of the doubly robust estimator because the weighted (iteratively reweighted) least squares estimator satisfies

$$\frac{1}{n} \sum_{i=1}^n \frac{R_i}{\pi_i} Y_i = \frac{1}{n} \sum_{i=1}^n \frac{R_i}{\pi_i} m^*(X_i).$$

With R_i/π_i being on average 1 at all levels of the background characteristics, it follows that $m^*(X_i)$ equals the outcome Y_i in expectation.

The proposed doubly robust estimator has the additional advantage of being more precise than the IPW estimator when both working models are correctly specified. This is because it extracts additional information from the imputation model (Robins, Rotnitzky & Zhao, 1995). This will emerge more clearly in the fourth section in the context of regression models with incomplete covariates. However, the doubly robust estimator is less precise than the imputation estimator $\hat{\mu}_{IMP}$ when the imputation model is correctly specified. This is the price to pay for this estimator which, in contrast to the imputation estimator, is unbiased in large samples when the outcome regression is mis-specified, provided that the probability of missing data is well modeled in function of background characteristics. The choice between doubly robust estimators and imputation estimators is thus mainly a tradeoff between bias and efficiency. Because the bias of an estimator is invisible, in contrast to its imprecision which is reflected through its standard error, the concern for bias usually trumps efficiency concerns and thus the doubly robust estimator may be the preferred one for routine use. However, the above reasoning is based on large sample arguments and may not provide a good picture of the finite sample behavior of the estimators, especially when the weights are extreme for some subjects. To gain some insight into this question, we report some small to moderate sample simulation experiments in the third section.

Alternatively, to reduce the chance of bias, one could envisage choosing flexible working models for the probability of missing data in the IPW (and doubly robust) estimator and for the expected outcome in the imputation (and doubly robust) estimator (Little & An, 2004). For the IPW and doubly robust estimator, increasing the flexibility of the former model not only reduces the potential for bias, but at the same time increases the precision of the estimator (so long as the model does not contain an excessive number of predictors) (Robins, Rotnitzky & Zhao, 1995). This flexibility in choosing a response model is an additional attraction of these estimators. In contrast, imputation estimators may lose substantial precision when flexible imputation models are adopted, and may then even become less precise than the IPW estimator in finite samples (see the third section). To overcome the potential for bias of the imputation estimator, without sacrificing too much precision, it has been suggested – with some good results – that one should reduce multiple background characteristics X_i into a single characteristic π_i which may then be modeled more flexibly (David, Little, Samuhel & Triest, 1983; Little & An, 2004).

3. Simulation study

In this section, we present results from two simulation experiments, each with sample sizes 200 and 500. The goal of these simulation studies is to understand better how the different estimators perform (a) when there is little overlap in the distribution of background characteristics between responders and nonresponders and, additionally, their association with outcome differs between these groups; and (b) when flexible working models are used. In both experiments, X and ε were generated from the standard normal

distribution. Next, Y was generated as $X^2 + \varepsilon$ in the first experiment and as $1 - \exp(X) + \varepsilon$ in the second experiment. Outcomes were then made missing by generating a binary missingness indicator R ($R = 0$ if missing and 1 otherwise) with missingness probability satisfying $\text{logit}\{P(R=1)\} = \alpha X$ and $\alpha=3$ ($\alpha=-3$). Under this simulation set-up, 50% of the data are missing and there is large separation between the covariate distribution of responders and nonresponders, so that the regression mean imputation estimator will suffer from extrapolation and the IPW estimator from extreme weights. These two data-generating mechanisms are illustrated, respectively, by the particular realizations shown in Figure 1 (top, left) for the first experiment and in Figure 2 (top, left) for the second experiment.

Figure 2 about here.

We used 1000 replications in each simulation experiment. In each replication, the following estimators were calculated. The IPW estimator in Equation 1 with weights estimated by fitting a k -th order (i.e., with predictors X^0, \dots, X^k) logistic regression model (IPW(k)), the regression mean imputation estimator in Equation 2 with a k -th order linear imputation model (RMI(k)), the doubly robust estimator with weights estimated by fitting a k -th order logistic regression model and mean imputations obtained from a l -th order linear imputation model (DR(k, l)), and the doubly robust estimator DR(4,4) with probabilities π_i truncated at 0.05 when they were estimated below 0.05 (DR trunc). Note that we only evaluate the impact of weight truncation for the doubly robust estimator because this estimator continues to be valid in the presence of misspecified weights (e.g., truncated weights) provided that the imputation model holds.

We used the nonparametric bootstrap (bias corrected and accelerated bootstrap percentile confidence intervals (Carpenter & Bithell, 2000)) for inference with all estimators. We used bootstrap standard error estimators because the usual calculations of standard errors and confidence intervals based on a sample average are invalid. For IPW estimators, this is because they ignore the imprecision of the weights, which were estimated using (logistic) regression models. For regression mean imputation estimators, this is because they ignore the imprecision of the fitted values $m(X_i)$, because these fitted values vary less than the real outcome and because they should not be considered as ‘real’ observations when calculating the standard error. For doubly robust estimators, this is true for both the above reasons.

Table 2 about here.

The results from the first and second simulation experiment are summarized in Tables 2 and 3, respectively. In both experiments, we evaluated the use of first to fourth order working models. Note that the first order imputation model is misspecified, but the first order response model is not, thus giving a relative advantage to the IPW estimator. The reason for our choice is that deviations from a linear model are difficult, or impossible, to detect in the imputation model as a result of separation in the covariate distribution of responders and nonresponders, while such deviations are detectable in the response model. We refer the reader to Bang and Robins (2005), Carpenter, Kenward and Vansteelandt (2006), Kang and Schafer (2008), Robins et al. (2008) and Vansteelandt, Rotnitzky and Robins (2007) for examinations on the impact of misspecification of the response model.

In the first experiment, the smallest variance occurs for the regression mean imputation estimator under a linear prediction model (RMI(1)). However, this estimator has a large bias as a result of extrapolation. This bias disappears upon choosing higher order imputation models, but at the expense of a greatly inflated variance when third or fourth order imputation models are used. This variance inflation is even more pronounced for the doubly robust estimator. In contrast, the IPW estimator handles the high dimensionality of the working model much better in the sense of having much smaller variance than the other estimators when higher order working models are used. This estimator is less biased than the regression mean imputation estimator RMI(1), but still significantly biased in finite samples. As can be seen in Figure 1 (top, right), this is because the separation in the covariate distribution between responders and nonresponders is so large that, with the considered sample sizes, no random sample from the study population can be ‘reconstructed’ using just the complete observations. As a result, confidence intervals based on this estimator do not attain the specified level of coverage. Even so, when flexible working models of order at least 3 are used, the smallest mean squared errors are seen for the IPW estimator, suggesting that it tends to end up closest to the true population mean.

Overall, as predicted by the theory, the best results are obtained for the regression mean imputation estimator RMI(2) under a correctly specified imputation model. In interpreting the results, one should note, however, (a) that the evidence for nonlinear imputation models may be weak when responders and nonresponders are very distinct in their prognostic factors X , and (b) that standard software for imputation usually does not accommodate nonlinear imputation models. With concern for bias, one could choose to

use flexible imputation models. However, this simulation experiment demonstrates that both imputation, and doubly robust, estimators cannot handle high-dimensional imputation models very well. The doubly robust estimator with truncated weights performs substantially better in terms of precision, but also does not succeed at approximating the mean squared error of the IPW estimator. The IPW estimator with flexible response model thus forms an attractive alternative: although prone to some finite-sample bias and undercoverage of confidence intervals in this setting, it tends to end up closest to the population mean.

The characteristic feature of the second simulation experiment is that all imputation models are now mis-specified. Similar results are obtained as in the first simulation experiment, with the exception that the doubly robust estimator with linear or second order imputation model and the doubly robust estimator with truncated weights now substantially outperform the other estimators in terms of mean squared error. This is because the misspecification of the linear model is insufficiently severe for the precision of this estimator to break down. In terms of coverage, the worst results are obtained via regression mean imputation with standard linear models and the best results with regression mean imputation or doubly robust estimation based on at least second order imputation models.

Table 3 about here.

4. Estimating regression models with incomplete outcome or covariate data

We first generalize the principle behind (doubly robust) IPW estimators. We then apply it to regression models with incomplete outcomes or incomplete covariates.

4.1. General inverse probability weighting estimators and doubly robust estimators

Virtually all standard estimators $\hat{\beta}$ of statistical parameters β are obtained as the solution to a so-called unbiased estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^n U_i(\hat{\beta}), \quad (4)$$

where $U_i(\beta)$ is a function of the observed data, which is unbiased in the sense that it has mean zero when evaluated at the true parameter value β^* . For instance, maximum likelihood estimators are obtained by solving Equation 4 with $U_i(\beta)$ the maximum likelihood score function. Likewise, the sample average is obtained by solving Equation 4 with $U_i(\beta) = Y_i - \beta$. Estimators obtained by solving such unbiased estimating equations are themselves unbiased for the true parameter value β^* when the sample size is sufficiently large (and weak regularity conditions hold). Intuitively, this follows from $U_i(\beta)$ having mean zero when evaluated at the true parameter value β^* and thus the sample analog, Equation 4, attaining zero at values $\hat{\beta}$ close to β^* .

When the data for some subjects are incomplete, the full estimating equation cannot be solved because the contribution of these subjects to the estimating equation is unknown. Evaluating the sample average in Equation 4 only for responders' data may no longer yield an unbiased estimating equation when missingness is selective, so the remaining subjects with no missing data are unrepresentative. Following the second section, this can be corrected for by weighting each observed contribution inversely by

the probability π_i of having completely observed data (given all data required to evaluate $U_i(\beta)$ and possibly additional background characteristics):

$$0 = \frac{\sum_{i=1}^n R_i U_i(\hat{\beta}) / \pi_i}{\sum_{i=1}^n R_i / \pi_i}. \quad (5)$$

The solution $\hat{\beta}$ to Equation 5 is also called an IPW estimator and remains unbiased in large samples because R_i/π_i is 1 on average so that Equations 4 and 5 are on average the same.

The IPW estimator obtained by solving Equation 5 can be very imprecise because it merely extracts information from responders, thus ignoring the partial information that may be available for nonresponders. In contrast, doubly robust estimators extract additional information from nonresponders and have the additional advantage of being valid when either one of two working models hold. In the spirit of the second section, they can be obtained by solving the predicted estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^n m_i(\hat{\beta}), \quad (6)$$

where $m_i(\beta)$ is the expected value of $U_i(\beta)$ calculated from the observed data using weights R_i/π_i : this can be obtained from a weighted least squares regression of $U_i(\beta)$ on the observed data, with weights R_i/π_i . This is illustrated in the following sections for the case of regression models with an incomplete outcome and incomplete covariate, respectively.

4.2. Regression models with missing outcome values, but fully observed covariates

Suppose that we have a study in which it is intended to collect outcome measurements Y_1, \dots, Y_n (e.g., income) and covariate measurements Z_1, \dots, Z_n (e.g., education) on a random sample of n subjects, in order to learn about the association between outcome and covariate. In particular, we may be interested in fitting a generalized linear model

$$E(Y) = g(\beta Z), \quad (7)$$

where $g(\cdot)$ is an (inverse) canonical link function (e.g., the identity link for a normally distributed outcome or the inverse logit link for a binary outcome) and where the covariate vector Z includes 1 to allow for an intercept. When the outcomes are missing for some subjects, then restricting the analysis to responders will not introduce bias when missingness is solely related to the covariate measurements Z , but may introduce bias when it is additionally related to prognostic factors X (e.g., social class) of the outcome that are not contained in Z . Although additionally including these prognostic factors in the regression model accommodates this, we may have good reasons not to do so. For example, this would be the case when X is affected by Z so that adjustment for X distorts the association between outcome and covariate Z (Rosenbaum, 1984). This is also the case when, for instance, the model of interest is a logistic regression model. Indeed, due to noncollapsibility of the odds ratio (Greenland, Robins & Pearl, 1999), the additional adjustment for X then changes the magnitude and interpretation of the other parameters in the model, even if X has no residual association with Y . In these cases, one may account for missingness being selective following the earlier methods. Specifically, by reweighting each responder's contribution, $Z_i\{Y_i - g(\beta Z_i)\}$, the score equation becomes

$$0 = \sum_{i=1}^n R_i Z_i \{Y_i - g(\hat{\beta} Z_i)\} / \pi_i,$$

valid estimators $\hat{\beta}$ are obtained provided that the probabilities π_i of complete outcome data are well modeled as a function of the covariates Z_i and the additional background characteristics X_i . In practice, IPW estimates are thus obtained using the following two-stage procedure:

1. Fit a logistic regression model for the probability of observing Y_i (i.e., $R_i = 1$) as a function of prognostic factors Z_i and X_i . Let π_i denote the fitted value for subject i .
2. Fit the generalized linear model of interest, i.e. Equation 7, to the responders, using weighted regression with weights $1/\pi_i$.

It also follows from the previous section that a doubly robust estimator can be obtained by fitting the model of interest, Equation 7, with the outcome for responders and nonresponders substituted by a fitted value. The latter is derived from a regression model for the outcome as a function of both X and Z ,

$$E(Y) = g(\gamma_z Z + \gamma_x X), \quad (8)$$

which is fitted to the responders, weighting each subject's contribution by the reciprocal of the fitted probability of complete data, i.e., $1/\pi_i$. Note that the model in Equation 8 is only used for predicting the outcome in the nonresponders, unlike the model of interest in Equation 7. In practice, doubly robust estimates are thus obtained by replacing step 2 in the above procedure by the following two steps:

2. Fit the generalized linear model in Equation 8 to the responders using weighted regression with weights $1/\pi_i$ and let $m^*(Z_i, X_i)$ denote the fitted value for subject i .
3. Fit the generalized linear model of interest in Equation 7 upon substituting the outcome of responders and nonresponders with $m^*(Z_i, X_i)$.

In large samples, the resulting estimator is at least as precise as the IPW estimator when the imputation model in Equation 8 is correctly specified. In addition, it is unbiased in large samples if either the response model or the imputation model is correctly specified.

4.3. Regression models with missingness in a single covariate

Suppose now that we have a study in which it is intended to collect outcome measurements Y_1, \dots, Y_n and covariate measurements $(Z_1, M_1), \dots, (Z_n, M_n)$ on a random sample of n subjects, but that the (possibly multivariate) covariate value M_i is missing for some subjects. Fitting the linear regression model

$$E(Y) = \beta_z Z + \beta_m M \quad (9)$$

to the responders may then introduce bias when missingness is possibly associated with the outcome Y or with prognostic factors X of outcome and covariate M . Including these prognostic factors in the regression model does not accommodate this problem when missingness is associated with the outcome. Instead, we may re-weight each responder's contribution to the standard normal equations for the model of interest by the reciprocal of the fitted probability π_i of complete covariate data, calculated as a function of the outcome Y_i , completely observed covariates Z_i and the additional background characteristics X_i . Note that, when the outcome is observed, it needs to be included in the model for the weights. This is possible with standard software using the following two-stage approach:

1. Fit a logistic regression model for the probability of observing X_i (i.e., $R_i = 1$) as a function of prognostic factors Y_i , Z_i and X_i . Let π_i denote the fitted value for subject i .

2. Fit the linear model in Equation 9 to the responders using weighted regression with weights $1/\pi_i$.

Because this approach ignores the partial information on Y_i and Z_i that is available on nonresponders, estimates may be very imprecise. It is therefore more attractive to calculate a doubly robust estimator following the methods that were introduced earlier. Below, we propose a novel, iterative procedure for obtaining such estimator on the basis of standard software routines:

1. Calculate expected values for M and M^2 as a function of Y , Z and X by postulating (separate) models for them and then fitting these models to the responders' data, weighting each subject's contribution by the reciprocal of the fitted probability of complete data. For instance, we may postulate a linear model for M

$$E(M) = \gamma_z Z + \gamma_y Y + \gamma_x X,$$

and, assuming that M has a constant variance, choose

$$E(M^2) = (\gamma_z Z + \gamma_y Y + \gamma_x X)^2 + \gamma.$$

Note that, having fitted the model for $E(M)$, the model for $E(M^2)$ contains just one unknown parameter and $(\gamma_z Z + \gamma_y Y + \gamma_x X)^2$ can be treated as an offset term.

2. Starting from the IPW estimates as initial estimates of β_z and β_m in Equation 9, repeat the following steps until the estimates for β_z and β_m converge:

- a. Fit model

$$E\{Y - \beta_m E(M)\} = \beta_z Z,$$

to all subjects with β_m evaluated at the current estimate, and obtain an updated estimate of β_z .

b. Fit model

$$E\{(Y - \beta_z Z)E(M)/E(M^2)^{1/2}\} = \beta_m E(M^2)^{1/2},$$

to all subjects with β_z evaluated at the current estimate, and obtain an updated estimate of β_m .

The resulting estimators for β_z and β_m which are obtained upon convergence of the algorithm, are again at least as precise as the IPW estimator when the sample size is sufficiently large and the regression models for M and M^2 are not too grossly misspecified. This is mainly because this doubly robust estimator additionally extracts information from the responders by substituting the missing covariate values with their expectation. In addition, as shown in the Appendix, this estimator is unbiased in large samples if either the response model or the two imputation models for M and M^2 are correctly specified. Note that this algorithm sides-steps the problem of non-monotone non-response, by grouping all the variables with missing data in ' M '.

5. Discussion

In this article, we have tried to give an intuitive explanation of the use of IPW and doubly robust estimators for the mean of an incomplete outcome and for regression parameters in generalized linear models with incomplete outcome or covariate data, emphasizing how to obtain these estimators via standard software. In addition, we have demonstrated the general principle which underpins these estimators in other contexts. In particular, (doubly robust) IPW estimators have been developed for handling drop-out or attrition (see e.g., Robins, Rotnitzky & Zhao, 1994, 1995) and intermittent missingness (Lin, Scharfstein & Rosenheck, 2004; Vansteelandt, Rotnitzky & Robins, 2007) in

longitudinal studies, for censoring adjustment in survival studies (Rotnitzky & Robins, 2005) and for handling missing not at random data (Scharfstein, Rotnitzky & Robins, 1999; Vansteelandt, Rotnitzky & Robins, 2007). So far, they have not been developed to handle general missingness patterns on multiple covariates. Work by the authors is ongoing to address these more general cases.

In line with other results in the literature, our simulations indicate that among the estimators considered, regression mean/multiple imputation yields the most precise estimators, provided that the imputation models are correctly specified, but not overspecified. This is because it is in effect an approximation to maximum likelihood. In some settings, the regression mean/multiple imputation may be tricky, however. First, because the imputation model must be sufficiently flexible with respect to the analysis model (Kenward & Carpenter, 2007). For instance, when the outcome model of interest involves interactions between an incomplete and complete covariate, the imputation model must be sufficiently rich not to a priori exclude such interactions. This is often difficult in practice. Similar difficulties occur when the model is nonlinear or when the data are clustered (Kenward & Carpenter, 2007). Second, the imputation model may be difficult to choose when responders and nonresponders are very dissimilar in terms of background characteristics because any statistical imputation model will be forced to make extrapolations beyond the range of the observed data. By avoiding imputation models altogether, IPW estimators do not suffer these problems. Similarly, doubly robust IPW estimators, although they rely on imputation models, seek to minimize the impact of these issues. Indeed, they incorporate the information in the imputation model in order to bolster their efficiency relative to IPW estimators, but do so in such a way as to minimize

the bias that occurs if the imputation model is mis-specified (provided that the response probabilities are well modeled).

The simulation studies in the third section suggest that the precision of multiple imputation estimators and doubly robust estimators can deteriorate quickly with the increasing complexity of the working models. As predicted by theory (Robins, Rotnitzky & Zhao, 1994), this is not the case for IPW estimators. These tend to be more precise and less biased with increasing complexity of the missingness model, but are prone to some finite-sample bias and undercoverage of confidence intervals when missingness is very selective (unless the sample size is large). In view of this, and because appropriate strategies for selecting imputation and response models have received relatively little attention, with few exceptions (Brookhart & van der Laan, 2006), we recommend trying various working models and estimation strategies for critical analyses. Based on the results in this article, we conjecture that model selection strategies may succeed better at identifying the true imputation model when inverse probability weighting is used to fit the imputation model, as in the doubly robust estimator. This is because estimation then focuses also on regions of the covariate space where the nonresponders are located. This is suggested by Figure 1 (bottom right panel), where the nonlinear pattern becomes more clearly apparent after inverse weighting. In future work, we will investigate whether, as anticipated, the benefits of doubly robust estimators become more pronounced in combination with careful model selection strategies. Further study is also warranted on how to best assess the standard errors of (doubly robust) IPW estimators in the presence of extreme weights. In this article, we have used the bootstrap for simplicity, but the bootstrap might not perform well when some individuals have large weights and get

oversampled in some of the bootstrap samples. As an alternative, closed-form 'sandwich' estimators have been proposed for the standard errors of (doubly robust) IPW estimators. These require programming and might not perform well in the presence of large weights because they are based on large sample approximations which may become poor when some individuals have large weights.

Acknowledgement

The authors are grateful to the guest editor and two referees for thorough and very helpful comments. They acknowledge support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy). James Carpenter and Mike Kenward are partially supported by a grant from the Medical Research Council, U.K. G0600599.

References

Bang, H., & Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference Models. *Biometrics*, 61, 692-972.

Brookhart, M.A., & van der Laan, M.J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics and Data Analysis*, 50, 475-498.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141-1164.

Carpenter, J., Kenward, M., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation. *Statistics in Society*, 169, 571-584.

David, M., Little, R.J.A., Samuël, M.E., & Triest, R.K. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 168-173.

Davidian, M., Tsiatis, A.A., & Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science*, 20, 261-301.

Greenland, S., Robins, J.M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14, 29-46.

Horvitz, D.G., & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kang, J.D.Y., & Schafer, J.L. (2008). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22, 523-539.

Kenward, M.G., & Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16, 199-218.

Lin, H.Q., Scharfstein, D.O., & Rosenheck, R.A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society – Series B*, 66, 791-813.

Little, R., & An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14, 949-968.

Murray, G.D., & Findlay, J.G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine*, 7, 941-946.

Robins, J.M., Rotnitzky, A., & Zhao, L.-P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

Robins, J.M., Rotnitzky, A., & Zhao, L.-P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

Robins, J.M., & Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87, 113-124.

Robins, J.M., & Rotnitzky, A. (2001). Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer", *Statistica Sinica*, 11, 920-936.

Robins, J.M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2008). Performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statistical Science*, 22, 544-559.

Rosenbaum, P.R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society Series A*, 147, 656-666.

Rotnitzky, A.G., & Robins, J. (2005). *Inverse Probability Weighted in Survival Analysis*. The Encyclopedia of Biostatistics. Vol 4. pp. 2619-2625. Second Edition. Edited by P. Armitage and T. Colton.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons, New York.

Scharfstein, D.O., Rotnitzky, A., & Robins, J.M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association*, 94, 1096-1120.

Tan, Z. (2008). Understanding OR, PS, and DR. *Statistical Science*, 22, 560-568.

Vansteelandt, S., Rotnitzky, A., & Robins, J.M. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94, 841-860.

Appendix

In this Appendix, we show that the estimators proposed in the section on incomplete covariates satisfy the double robustness property. First, note that by construction, the fitted values are obtained from a weighted least squares regression of models $E(M) = \gamma_z Z + \gamma_y Y + \gamma_x X$ and $E(M^2) = (\gamma_z Z + \gamma_y Y + \gamma_x X)^2 + \gamma$, with weights $1/\pi_i$. When these models are correctly specified, the obtained fitted values will be consistent, even if the weights are misspecified, by the fact that the models are conditional on (X, Y, Z) and the weights are functions of (X, Y, Z) . In addition, these fitted values satisfy

$$\begin{aligned} \sum_{i=1}^n R_i Z_i M_i / \pi_i &= \sum_{i=1}^n R_i Z_i E(M_i) / \pi_i \\ \sum_{i=1}^n R_i Y_i M_i / \pi_i &= \sum_{i=1}^n R_i Y_i E(M_i) / \pi_i \\ \sum_{i=1}^n R_i M_i^2 / \pi_i &= \sum_{i=1}^n R_i E(M_i^2) / \pi_i. \end{aligned} \tag{A1}$$

Second, note that the estimates for β_z and β_m satisfy $\sum_{i=1}^n Z_i \{Y_i - \beta_z Z_i - \beta_m E(M_i)\} = 0$ and

$\sum_{i=1}^n E(M_i) \{Y_i - \beta_z Z_i\} - \beta_m E(M_i^2) = 0$ and thus solve

$$\sum_{i=1}^n E\{U_i(\beta_z, \beta_m) | Z_i, Y_i\} = 0, \quad (\text{A2})$$

with $U_i(\beta_z, \beta_m) = (Z_i' \ M_i')'(Y - \beta_z Z - \beta_m M)$. The double robustness is now immediate because, if the imputation models for M and M^2 are correctly specified, then Equation A2 is an unbiased estimator of the population mean equations and thus an unbiased estimating equation. In contrast, when the probabilities of complete covariate data are correctly specified, then because the equalities given in Equation A1 imply

$\sum_{i=1}^n \frac{R_i}{\pi_i} [U_i(\beta_z, \beta_m) - E\{U_i(\beta_z, \beta_m) | Z_i, Y_i\}] = 0$ and because R_i/π_i is on average 1, we have

that Equation A2 equals $\sum_{i=1}^n U_i(\beta_z, \beta_m)$ in expectation.

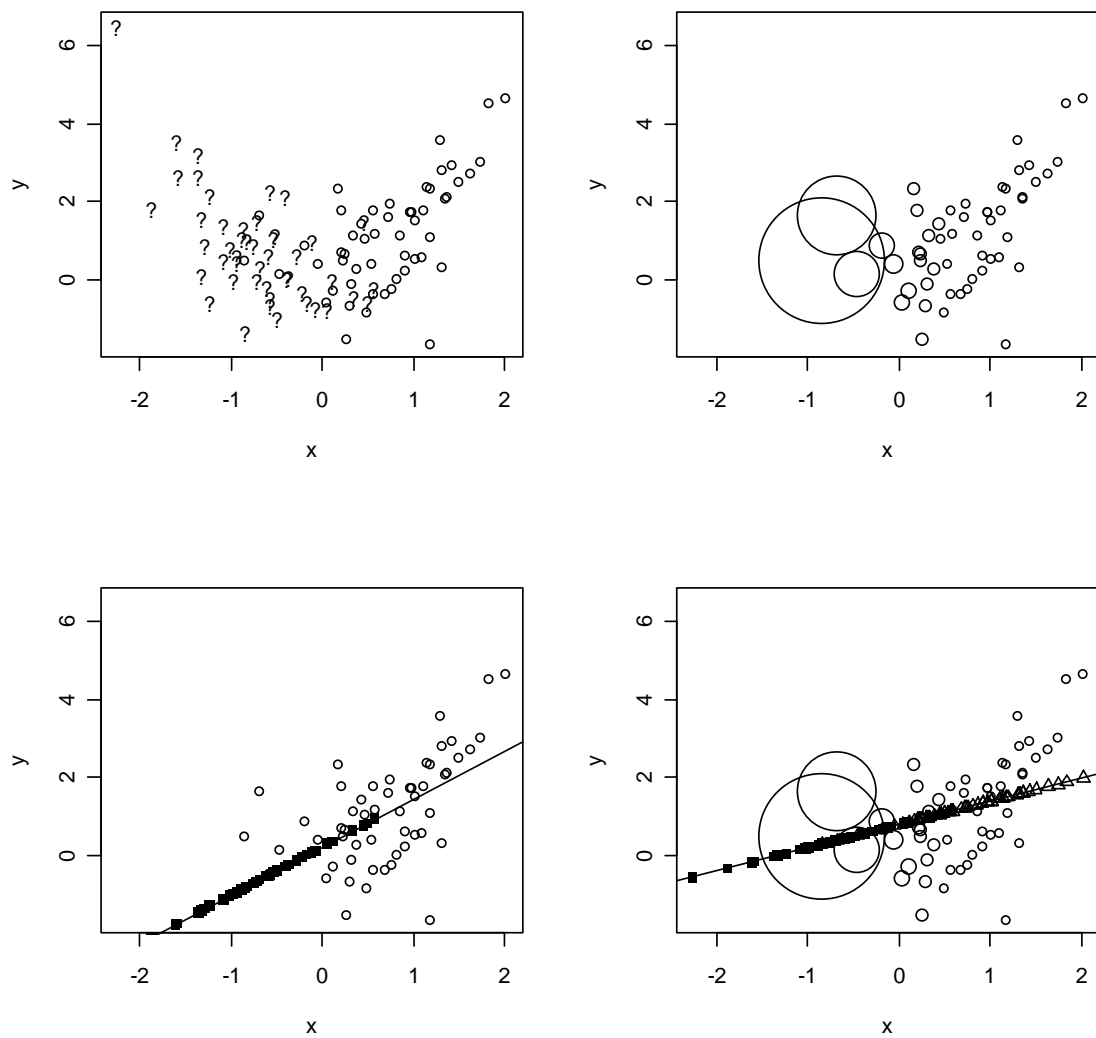


Figure 1. Simulation experiment 1: Observed (small circles) and missing (?) observations (top, left); Inverse probability weighted observations (top, right); Observations for responders and OLS predictions for nonresponders (bottom, left); Observations for responders and weighted least squares predictions for nonresponders (bottom, right).

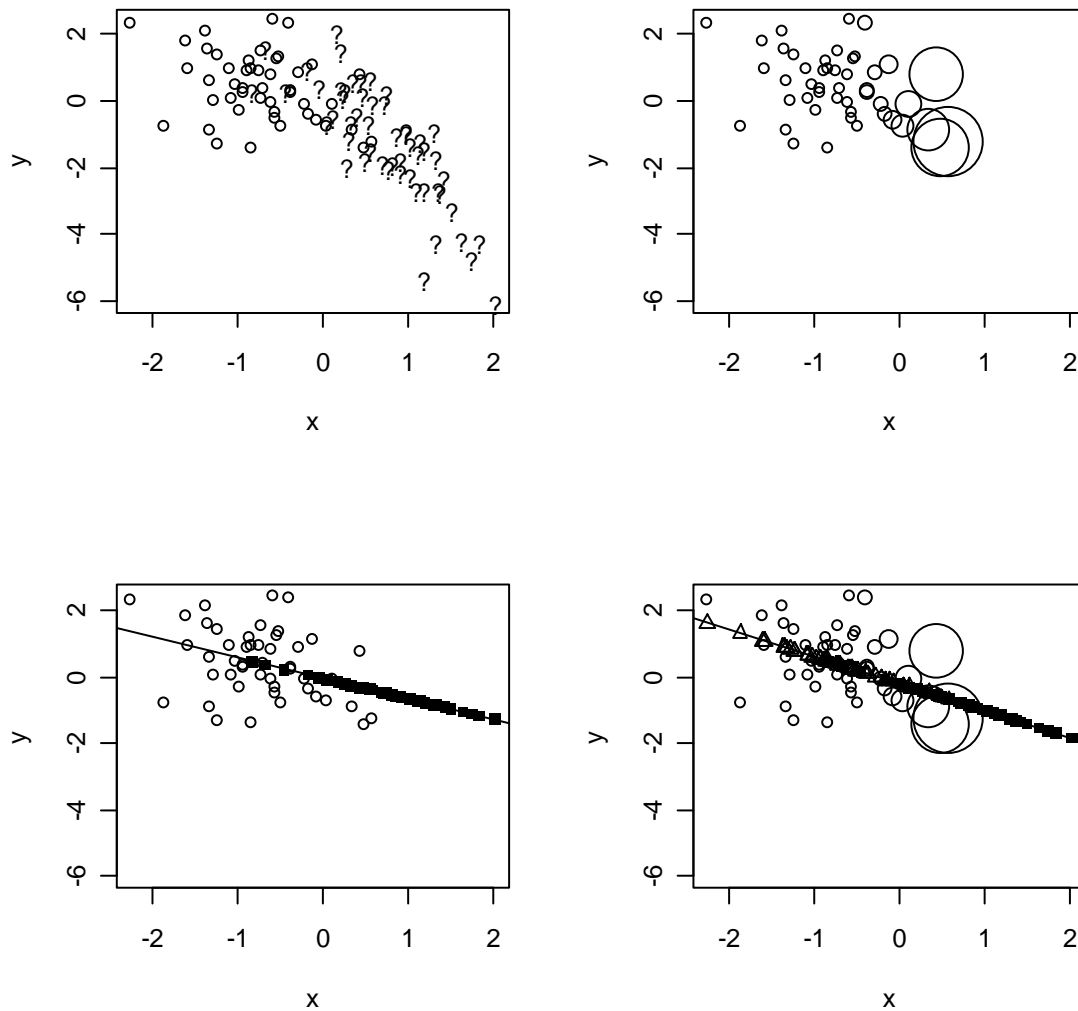


Figure 2. Simulation experiment 2: Observed (small circles) and missing (?) observations (top, left); Inverse probability weighted observations (top, right); Observations for responders and OLS predictions for nonresponders (bottom, left); Observations for responders and weighted least squares predictions for nonresponders (bottom, right).

Table 1: Toy example.

X_i	A	A	A	B	B	B	C	C	C
Y_i	1	1	1	2	2	2	3	3	3
R_i	1	0	0	1	1	1	1	1	0
π_i	1/3			1	1	1	2/3	2/3	
R_i/π_i	3	0	0	1	1	1	1.5	1.5	0

Table 2: Median bias and empirical variance of the estimates, empirical coverage probability and average length of 95% confidence intervals, and mean squared error (MSE) in simulation experiment 1 for sample sizes 200 and 500.

N	Estimator	Median bias	Empirical variance	Coverage 95% CI	Average length 95% CI	MSE
200	IPW(1)	-0.24	0.12	0.64	0.67	0.15
	DR(1,1)	-0.41	0.16	0.53	0.93	0.30
	RMI(1)	-1.11	0.052	0.00	0.84	1.3
	IPW(2)	-0.24	0.067	0.64	0.76	0.11
	DR(2,2)	-0.012	0.097	0.93	1.98	0.097
	RMI(2)	0.0055	0.055	0.94	1.78	0.055
	IPW(3)	-0.24	0.060	0.67	0.76	0.10
	DR(3,3)	-0.0068	0.26	0.93	4.06	0.26
	RMI(3)	0.0013	0.17	0.94	1.77	0.17
	IPW(4)	-0.24	0.11	0.72	0.99	0.17
	DR(4,4)	-0.0075	$3.7 \cdot 10^7$	0.93	891	$3.7 \cdot 10^7$
	DR(4,1)	-0.41	2.3	0.66	1.31	2.4
	DR(1,4)	-0.016	1.2	1.00	5.54	1.2
	DR trunc	-0.44	0.095	0.51	0.91	0.31
	RMI(4)	0.0018	0.88	0.98	2.86	0.88
500	IPW(1)	-0.20	0.12	0.58	0.50	0.14
	DR(1,1)	-0.31	0.097	0.56	0.73	0.17
	RMI(1)	-1.13	0.021	0.00	0.52	1.3
	IPW(2)	-0.20	0.063	0.58	0.64	0.092
	DR(2,2)	-0.0038	0.038	0.94	1.03	0.038
	RMI(2)	0.00080	0.022	0.97	0.89	0.022
	IPW(3)	-0.19	0.039	0.60	0.65	0.069
	DR(3,3)	0.0021	0.073	0.94	1.03	0.073
	RMI(3)	0.0027	0.043	0.96	0.89	0.043
	IPW(4)	-0.20	0.057	0.62	0.69	0.089
	DR(4,4)	-0.0017	$6.6 \cdot 10^7$	0.95	3.52	$6.6 \cdot 10^7$
	DR(4,1)	-0.29	0.14	0.60	0.85	0.22
	DR(1,4)	0.00024	0.23	1.00	2.03	0.23
	DR trunc	-0.41	0.035	0.41	0.66	0.20
	RMI(4)	-0.0045	0.14	0.96	1.34	0.14

Table 3: Median bias and empirical variance of the estimates, empirical coverage probability and average length of 95% confidence intervals, and mean squared error (MSE) in simulation experiment 2 for sample sizes 200 and 500.

N	Estimator	Median Bias	Empirical variance	Coverage 95% CI	Average length 95% CI	MSE
200	IPW(1)	0.52	0.34	0.47	0.85	0.48
	DR(1,1)	0.36	0.086	0.55	0.77	0.19
	RMI(1)	0.61	0.026	0.040	0.60	0.40
	IPW(2)	0.52	0.12	0.42	1.01	0.32
	DR(2,2)	0.22	0.097	0.92	2.05	0.14
	RMI(2)	0.31	0.057	0.92	1.79	0.15
	IPW(3)	0.51	0.11	0.46	1.04	0.32
	DR(3,3)	0.14	0.26	0.93	2.06	0.27
	RMI(3)	0.15	0.18	0.92	1.79	0.20
	IPW(4)	0.49	0.13	0.66	1.19	0.31
	DR(4,4)	0.074	$5.1 \cdot 10^7$	0.94	751	$5.1 \cdot 10^7$
	DR(4,1)	0.36	1.0	0.68	1.07	1.2
	DR(1,4)	0.56	1.9	0.97	3.84	2.4
	DR trunc	0.38	0.057	0.53	0.78	0.20
	RMI(4)	0.083	0.73	0.97	2.81	0.73
500	IPW(1)	0.44	0.50	0.51	0.74	0.57
	DR(1,1)	0.31	0.068	0.49	0.57	0.14
	RMI(1)	0.61	0.011	0.00	0.37	0.39
	IPW(2)	0.45	0.20	0.42	0.99	0.31
	DR(2,2)	0.19	0.049	0.87	1.04	0.083
	RMI(2)	0.31	0.023	0.87	0.89	0.12
	IPW(3)	0.43	0.11	0.46	1.01	0.23
	DR(3,3)	0.11	0.088	0.88	1.04	0.10
	RMI(3)	0.16	0.047	0.88	0.89	0.072
	IPW(4)	0.41	0.14	0.55	0.86	0.24
	DR(4,4)	0.059	$1.5 \cdot 10^5$	0.95	3.48	$1.5 \cdot 10^5$
	DR(4,1)	0.30	0.12	0.56	0.67	0.20
	DR(1,4)	0.057	0.27	0.95	1.91	0.27
	DR trunc	0.36	0.026	0.38	0.54	0.15
	RMI(4)	0.065	0.14	0.95	1.36	0.15